# MISSENSE MEANDERINGS IN SEQUENCE SPACE: A BIOPHYSICAL VIEW OF PROTEIN EVOLUTION

*Mark A. DePristo, Daniel M. Weinreich and Daniel L. Hartl*

Abstract | Proteins are finicky molecules; they are barely stable and are prone to aggregate, but they must function in a crowded environment that is full of degradative enzymes bent on their destruction. It is no surprise that many common diseases are due to missense mutations that affect protein stability and aggregation. Here we review the literature on biophysics as it relates to molecular evolution, focusing on how protein stability and aggregation affect organismal fitness. We then advance a biophysical model of protein evolution that helps us to understand phenomena that range from the dynamics of molecular adaptation to the clock-like rate of protein evolution.

FIXATION
A mutation that has achieved a frequency of 100% in a natural population.

ADAPTIVE EVOLUTION
A genetic change that results in increased fitness.

FITNESS
A measure of the capacity of an organism to survive and reproduce.

*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.*
*Correspondence to M.A.D. e-mail: mark_depristo@harvard.edu*

doi:10.1038/nrg1672
Published online 2 August 2005

Taken as a whole, recent findings from biochemistry and evolutionary biology indicate that our understanding of protein evolution is incomplete, if not fundamentally flawed. The neutral theory of molecular evolution[1], which states that all mutations that reach FIXATION in a population are selectively neutral, appeals to evolutionary geneticists in part because it can account for the approximately constant rate of protein evolution. However, its premise that most missense mutations are selectively neutral has been systematically rejected by protein biochemists, who recognize instead that almost all missense mutations have large biophysical effects[2]. Indeed, nucleotide sequence analyses have uncovered pervasive positive selection for amino-acid replacements[3–5].

Another important challenge to evolutionary theory, which emphasizes the independent and additive effects of mutations, arises from studies of compensatory evolution. Here the deleterious effects of mutations are rapidly and effectively compensated by conditionally beneficial mutations. Compensatory mutations often occur in the same gene as the initial deleterious mutation, are common in ADAPTIVE EVOLUTION[6–8] and have an important role in many human diseases[9]. There are currently no models that reconcile the constant rate of protein evolution with the biochemical reality that missense mutations have large, context-dependent effects and that few, if any, are selectively neutral.

There is a growing appreciation of the role that the biophysical properties of protein stability, aggregation and degradation have in FITNESS and disease[10] (TABLE 1). Moreover, these properties have been identified as significant factors in many cases of adaptive[8,11,12] and compensatory evolution[13–15]. These properties — and not function — seem to be the forces driving much of protein evolution.

Here we review the literature on biophysics as it relates to molecular evolution, with a particular focus on how missense mutations affect protein stability and aggregation. We then develop a biophysical model of protein evolution that helps to explain such diverse phenomena as compensatory mutation, the dynamics of molecular adaptation and the rate of protein evolution. Throughout this review, we bring together the fields of protein biophysics and molecular evolution by highlighting the shared questions, complementary techniques and important results concerning protein evolution that have come from both fields.

## Basic principles of protein biophysics

***Folding and stability.*** Decades of experimental and theoretical work have provided a detailed mechanistic

Table 1 | **Human diseases caused by defects in protein folding, stability and aggregation**

| Disease | Protein affected | Description | References |
|---|---|---|---|
| Cystic fibrosis | Cystic fibrosis transmembrane conductance regulator (CFTR) | The ΔPhe508 mutant has wild-type activity, but impaired folding in the endoplasmic reticulum leads to degradation. | 97 |
| α1 Antitrypsin deficiency | α1 Antitrypsin (also known as SERPINA1) | 80% of Glu342Lys mutants misfold and are degraded. Pathology is due to aggregation in patients with a reduced degradation rate. | 97 |
| SCAD deficiency | Short-chain acyl-CoA dehydrogenase (SCAD) | Impaired folding of Arg22Trp mutants leads to rapid degradation. | 98 |
| Alzheimer disease | Presenilin, γ-secretase | Mutations cause incorrect cleavage by the γ-secretase protease to produce the amyloid β-peptide; this aggregates into extracellular amyloid plaques. | 99,100 |
| Parkinson disease | α-Synuclein | Oxidative damage causes misfolding and aggregation. Hereditary forms are linked to deficiency in ubiquitin-mediated degradation. | 101 |
| Huntington disease | Huntingtin | CAG expansions in the Huntingtin gene lead to an abundance of polyglutamine fragments that aggregate and associate non-specifically with other cellular proteins. | 101,102 |
| Sickle cell anaemia | Haemoglobin | The Glu6Val mutation leads to aggregation in red blood cells. | 103 |

understanding of the forces that govern the folding and misfolding of two-state proteins[16]. Such proteins undergo a rapid transition from an unstructured, random conformation (the unfolded state) to a unique conformation called the native state, in which the protein carries out its function. Individual molecules are continuously unfolding and refolding, a phenomenon termed breathing[16]. The observations and arguments presented here, although clearest for two-state proteins, can be generalized to more complex, multi-state proteins. Nevertheless, most single-domain and short proteins (<110 amino acids) follow two-state folding kinetics.

The thermodynamic stability of a protein ($\Delta G$) is a measure of the ratio of folded to unfolded molecules, independent of the pathway followed during folding (see supplementary information S1 (box) for details of how $\Delta G$ is calculated)[17]. Low stability results in a large pool of unfolded molecules, whereas high stability results in more folded molecules, a more rigid structure and increased resistance to denaturants, such as urea and temperature. Proteins are only marginally stable, with $\Delta G$ values that seem to be constrained to between –3 and –10 kcal mol$^{-1}$. This range is based on observed $\Delta G$ values among unrelated proteins[17,18], although no systematic study has been undertaken. To put these values in context, the energy of a single hydrogen bond is 2–10 kcal mol$^{-1}$ (REF. 19).

***Aggregation.*** In addition to being unstable, proteins often have problems folding correctly and sometimes become trapped in misfolded conformations or form aggregates. These aggregates are non-functional, insoluble, cytotoxic and contain many molecules[10,20]. The misfolding occurs because the interactions that stabilize the native state are at least as favourable in aggregates. As it is an associative process, the rate of aggregation ($k_{agg}$) increases nonlinearly with increasing protein concentration[21]. This is exacerbated by the crowded intracellular environment, which favours association[21].

Although it seems that all proteins are able to aggregate, they differ substantially in their intrinsic propensity to do so under physiological conditions[22]. Recent studies have shown that this propensity is largely determined by the protein's stability, charge and tendency to form β SHEETS[23]. Although aggregation rate is related to stability, the two are not equivalent, as highlighted by mutations that affect aggregation independently of stability[20,24,25]. Importantly, aggregation seems to directly harm cells[26] — for example, in Huntington or Alzheimer disease (TABLE 1). The importance of proper folding is best highlighted by the pervasiveness of cellular mechanisms that discourage, isolate and eliminate aggregated proteins, such as CHAPERONES[27], INCLUSION BODIES[28] and the UBIQUITIN–PROTEASOME PATHWAY[29].

***Degradation.*** Cells continuously synthesize and degrade proteins. Because the degradation machinery operates selectively on partially or fully unfolded proteins, degradation rates are mainly determined by stability[30]. However, misfolded or abnormal proteins are also selectively targeted for degradation[31]. Moreover, rapid protein turnover is essential for regulation. Unsurprisingly, abnormal protein degradation is common in human genetic disease[32] (TABLE 1). For brevity, we do not discuss degradation further as a phenomenon that is independent of stability and aggregation.

***Constraints on protein folding and stability.*** Natural selection limits $\Delta G$ values to between –3 and –10 kcal mol$^{-1}$, implying a fitness penalty for proteins with stabilities that lie outside these boundaries. The lower limit is commonly appreciated and easy to

β SHEET
A secondary protein structure that has extensive, non-local hydrogen bonding.

CHAPERONES
A large class of cellular proteins that help other proteins to fold into their correct native conformation.

INCLUSION BODIES
Insoluble aggregates of misfolded proteins; inclusion bodies are common in prokaryotes.

UBIQUITIN–PROTEASOME PATHWAY
A eukaryotic degradation system in which ubiquitin molecules are attached to a target protein that is subsequently degraded by the proteasome complex.

understand: an unstable protein has a decreased EFFECTIVE CONCENTRATION owing to a large unfolded population that is rapidly degraded and/or aggregates[33]. The reason for the upper limit is more subtle and less well-studied, but has important implications. Despite the impression of rigidity from structural studies, proteins are in fact dynamic molecules, and their functions are critically dependent on mechanical flexibility[19,34,35]. Increasing stability results in a concomitant loss of flexibility and activity[36]. Moreover, highly stable proteins are protease-resistant and therefore difficult to regulate. In systems such as cell signalling, where removing a signal is as important as its activation, many proteins are actually NATIVELY UNFOLDED to ensure their rapid degradation[37].

Proteins must function reliably over a range of temperatures in POIKILOTHERMIC ORGANISMS. For example, *Escherichia coli* grows well at 21–49°C (REF. 38), corresponding to a $\Delta G$ range that spans 0.4 and 0.9 kcal mol$^{-1}$ for proteins with $\Delta G$ values of –5 and –10 kcal mol$^{-1}$ at 30°C, respectively. Improper folding, rather than loss of function, limits growth at extreme temperatures, as cell viability in these conditions can be recovered by overexpressing native folding chaperones or their EXTREMOPHILE analogues [38]. Consequently, proteins can commonly tolerate changes in $\Delta G$ that are on the order of ~1.0 kcal mol$^{-1}$ without significant loss of activity.

## Biophysical effects of mutation and selection

Stability and aggregation seem to have at least as great a role in protein evolution as in cellular and organismal function. Most missense mutations result in large perturbations of stability and aggregation. Consequently, such mutations might dominate the evolutionary dynamics of proteins.

***Mutational effects on protein stability and aggregation.*** A missense mutation can be either stabilizing or destabilizing. The effect of a mutation can be described by the difference in free energy between the wild-type and mutant forms of the protein, which is termed $\Delta\Delta G$ (see supplementary information S2 (figure)). Importantly, the effects of multiple mutations are approximately additive, although individual mutations show strong EPISTASIS for $\Delta G$ with a handful of other sites in the protein[39,40].

$\Delta\Delta G$ values are often of the same magnitude as $\Delta G$, as most single-residue mutations alter $\Delta G$ values by 0.5–5 kcal mol$^{-1}$ (REFS 14,33,41–47). Even the most conservative mutations at the most tolerant sites usually change $\Delta G$ by >0.1 kcal mol$^{-1}$ (REF. 41). Mutations that cause the greatest loss of stability — for example, those that introduce polar residues into the hydrophobic protein core — can destabilize a protein by >5 kcal mol$^{-1}$, often resulting in completely unfolded and therefore inactive proteins[33,41,48]. For example, two-thirds of missense mutations in the bacteriophage $\lambda$ Cro protein[49]; half the alanine mutations in the phage p22 Arc repressor[42]; and sixteen out of nineteen amino-acid mutations at a non-functional site in $\beta$-lactamase[48] significantly affect stability. The overwhelming conclusion from 20 years of mutational studies of protein stability is that most amino-acid replacements, at all sites in a protein, result in large effects on $\Delta G$ relative to the observed range of $\Delta G$ values themselves.

Aggregation rates are similarly sensitive to mutations. Destabilizing mutations result in increased aggregation rates owing simply to the larger pool of unfolded molecules[20,25]. In addition, a class of mutations has been identified that affect aggregation rate independently of changes in stability[15,22–25,28,50–52]. Even short sequence motifs can be AMYLOIDOGENIC, and their under-representation in sequence databases indicates that they are avoided by natural selection[53,54].

The large effect of mutations on stability and aggregation indicates that only a small number of missense mutations can be selectively neutral. Indeed, many single mutants should be significantly impaired by stability or aggregation defects, a prediction that is consistent with a survey of human genetic diseases (TABLE 1). Moreover, mutations that affect stability and aggregation, unlike those that affect function, are distributed across the protein molecule[33,41,47]. Overall, the existing evidence supports the general hypothesis that almost all mutations, at all sites in a protein, affect stability and aggregation. This is in stark contrast to mutations that affect function, which are generally restricted to a small number of specific catalytic residues.

***Pleiotropy among function, stability and aggregation.*** Recently there has been a growing appreciation that protein function, stability and aggregation are intrinsically linked, and this has important implications for the overall effects of mutation. Protein function depends on mechanical flexibility, which is linked to stability, so that increased stability results in a rigid molecule with reduced enzymatic activity[19,34–36,55]. At least in enzymes, functional residues are necessarily grouped together in space within active sites, and are often sequestered from water to provide a controlled reaction environment[56]. This organization is thermodynamically unfavourable because functional residues are generally polar or charged and therefore hydrophilic[56]. Consequently, functional residues are likely to be destabilizing. This was demonstrated by recent experiments in which the mutation of active-site residues to hydrophobic amino acids significantly increased stability, and concomitantly reduced activity[8,11,12]. The trade-off between activity and stability leads to evolutionary dynamics whereby functional adaptation results in a destabilized enzyme, requiring compensatory mutations to restore stability[8].

This demonstrates that mutations are pleiotropic at the biochemical level, and simultaneously affect stability, aggregation and activity. Although studies generally focus on just one phenotype, we suggest that random mutations will perturb all of these properties to some degree. Moreover, as the effects of individual mutations differ in the extent that they affect each of these properties, single amino-acid replacements that perturb a protein favourably along one dimension will probably be unfavourable along another. In particular, an adaptive change in function will demand a series of other fixation events to restore biochemical constraints.

EFFECTIVE CONCENTRATION
The concentration of functional molecules, as opposed to the total concentration of molecules.

NATIVELY UNFOLDED
Describes a class of proteins that are unfolded under physiological conditions.

POIKILOTHERMIC ORGANISM
An organism in which body temperature fluctuates with environmental temperature.

EXTREMOPHILE
An organism that thrives in environments that are inhospitable to most other organisms such as extreme heat (thermophiles), salinity (halophiles) and pressure (barophiles).

EPISTASIS
This occurs when the effect of a mutation varies with genetic background.

AMYLOIDOGENIC
Describes a protein that forms amyloid fibrils — a large, extended conformation that is adopted by many aggregated proteins. Amyloid fibrils are characteristic of several neurological disorders.

### Box 1 | A mathematical model for protein evolution

Fitness effects and the distribution of protein stabilities that follow mutation can be represented using a simple mathematical model.

**Relationship between stability and fitness**

In our model, organismal fitness, $W$, is related to the deviation from an optimal stability, $\Delta G_{opt}$, for a particular protein using a modified normal distribution (equation 1):

$$W(\Delta G) \propto \exp\left(-\left[\frac{\Delta G - \Delta G_{opt}}{\sigma_{\Delta G}}\right]^4\right) + c \qquad (1)$$

where $\Delta G$ is stability, $\sigma_{\Delta G}$ determines the breadth of the distribution, and $c$ ($0 \leq c \leq 1$) is the upper bound on the fitness cost to the organism owing to effects such as protein misfolding, degradation, aggregation and loss of regulation. $c$ reflects the exposure of the protein to natural selection that is due to its effect on organismal fitness, which would be high for essential proteins and low for non-essential proteins.

Biological considerations described in the main text indicate that $\Delta G_{opt}$ ranges between –4 and –8 kcal mol$^{-1}$ and $\sigma_{\Delta G}$ is approximately 2–3 kcal mol$^{-1}$, although these parameter values are likely to vary among proteins. A representative fitness function for an essential protein derived using this model is presented in supplementary information S3 (figure) (values of $c = 1$, $\Delta G_{opt} = -4$ kcal mol$^{-1}$ and $\sigma_{\Delta G} = 2.5$ kcal mol$^{-1}$ are used). Note that this equation above might be valid only when $\Delta\Delta G$ is small, as the fitness of extremely unstable or stable sequences is unlikely to be symmetrical, although natural selection will keep populations from these extremes. It should also be noted that $\Delta G_{opt}$ is an idealized biochemical property that might not be realized by any sequence in the population.

**The distribution of effects of mutation**

We model the distribution of mutational effects as a probability density function (PDF) that is based on the sum of two normal distributions with equal means, μ, and variances, $\sigma_{\Delta\Delta G}$, (equation 2):

$$\mathrm{PDF}(\Delta\Delta G) \propto k_1 \exp\left(-\left[\frac{(\Delta\Delta G - \mu)}{\sigma_{\Delta\Delta G}}\right]^2\right) + k_2 \exp\left(-\left[\frac{(\Delta\Delta G + \mu)}{\sigma_{\Delta\Delta G}}\right]^2\right) \qquad (2)$$

where the ratio $k_1/k_2$ reflects the relative abundance of destabilizing ($k_1$) compared with stabilizing ($k_2$) mutations. Again, biological considerations place μ at around 2 kcal mol$^{-1}$, $\sigma_{\Delta\Delta G}$ at around 1 kcal mol$^{-1}$, and the $k_1/k_2$ ratio at around 2, and these are the values used in supplementary information S3 (figure). These parameter values will also vary among proteins.

Consequently, treating stability, aggregation and activity as separable, independently varying parameters is inherently misleading.

***Selection for stability and aggregation.*** Proteins with altered stabilities and aggregation rates will frequently be generated by missense mutations at the many sites that affect these biophysical properties. Consequently, natural selection will 'see' a range of stabilities and aggregation rates, and will act on the favourable or unfavourable alternatives. Although mutants that affect these properties will usually be deleterious, in some cases an altered stability and aggregation rate can be highly advantageous[57]. Although we have said less about the effects of natural selection on aggregation than on stability, evolution seems to have avoided sequences that have a high aggregation propensity[10,53].

A particularly clear example of how natural selection acts on mutations that affect biophysical properties is provided by temperature adaptation[36,55]. Comparisons of ORTHOLOGOUS PROTEINS from species that live at a broad range of temperatures demonstrate that stability is tuned so that structural and functional characteristics are equilibrated with environmental temperature[36,55]. Heat-adapted proteins have increased thermostability to ensure proper folding at increased environmental temperatures[58,59]. Conversely, cold-adapted enzymes have reduced thermostability, which allows them not only to fold at low-temperatures, but also to preserve the flexibility essential for catalysis[34–36].

***Protein biophysics and compensatory mutation.*** As compensatory mutations mask the deleterious effects of other mutations[60], they are, by definition, conditionally beneficial, and therefore must be either deleterious or neutral on wild-type backgrounds[60–62]. They are commonly detected as mutations that restore protein activity[14,15,63–65] or organismal fitness[61,62,66–69]. The frequency of compensatory mutations in proteins is surprisingly high[6,7,9,62,66,69,70], with around 10–12 compensatory mutations for each deleterious mutation[7,62].

The large number of compensatory mutations implies that compensatory mechanisms must involve general biophysical properties such as stability and aggregation, which are determined by many residues, rather than mutations that affect the much smaller number of functional residues. The biochemical basis for many compensatory mutations has been explored in various systems, both by screening for mutations that restore the activity of engineered, deleterious mutations,

ORTHOLOGOUS PROTEINS
Proteins corresponding to genes that are related through speciation. By contrast, paralogous proteins are related by gene duplication.

and by examining the deleterious side-effects of advantageous, functional mutations. Many studies have inserted destabilizing and/or aggregation-promoting amino-acid replacements into proteins through genetic engineering and have screened for compensatory mutations[14,15,63–65]. Many second-site mutations have been isolated, and have been found to vary in terms of the residues affected[15,64,71]. Both direct biophysical measurements and computational modelling indicate that many pairs of primary and compensatory mutations act by first impairing and then restoring protein stability[13–15,64,65,72,73].

Because drug-resistance mutations often decrease organismal fitness[61,66–68,74], another common strategy to study compensatory mutations is to first screen for primary mutations that confer resistance, and subsequently screen for compensatory mutations that restore fitness[61,75]. The trade-off between resistance and fitness mirrors the trade-off between protein function, stability and aggregation. This probably reflects the loss of fitness that is due to the underlying destabilizing and aggregation-promoting effects of functional mutants. As these primary functional changes generally affect stability and aggregation, subsequent mutations that restore stability and reduce aggregation become highly favourable. The evolution of resistance to the antibiotic cefotaximine in *E. coli* β-lactamase demonstrates the dynamics of adaptation and compensation, as a catalytic mutation (Gly238Ser) is coupled with a stabilizing/anti-aggregation mutation (Met182Thr)[8,71].

Interestingly, a large number of individual second mutations can compensate for many initial deleterious mutations. These so-called 'global suppressors' can affect both stability and aggregation, apparently independently[15,64,65,71,73]. Protein-folding chaperones might be regarded simply as higher-level global suppressors, which unveil a kaleidoscope of latent genetic variation when they are inhibited[76,77]. The phenotypic effects of inhibiting chaperones result from uncovering previously suppressed folding, stability and aggregation effects[76,77]. Therefore, the inhibition of chaperones is similar in principle to removing intragenic suppressor mutations, differing only in the breadth of the effects revealed.

Compensatory mutations pose two challenges to traditional population genetic theory, which commonly assumes a near-absence of epistasis and the serial fixation of individual beneficial mutations that have small effects. First, many protein-coding genes have compensatory mutations for primary mutations that either do not seem to provide a selective advantage, and are therefore neutral, or are probably deleterious[9,70,73]. Therefore, in many cases it seems that the primary and compensatory mutations are individually deleterious but jointly neutral[78]. Second, even with compensation, mutants are often less fit than the wild type[74], but double-mutant genotypes are observed at a high frequency relative to revertants, even in environments that mask the benefits of the primary mutation[62,66]. The knowledge that has accumulated about the biophysical effects of mutation on protein stability and function allows us to develop a model of protein evolution that addresses these issues.

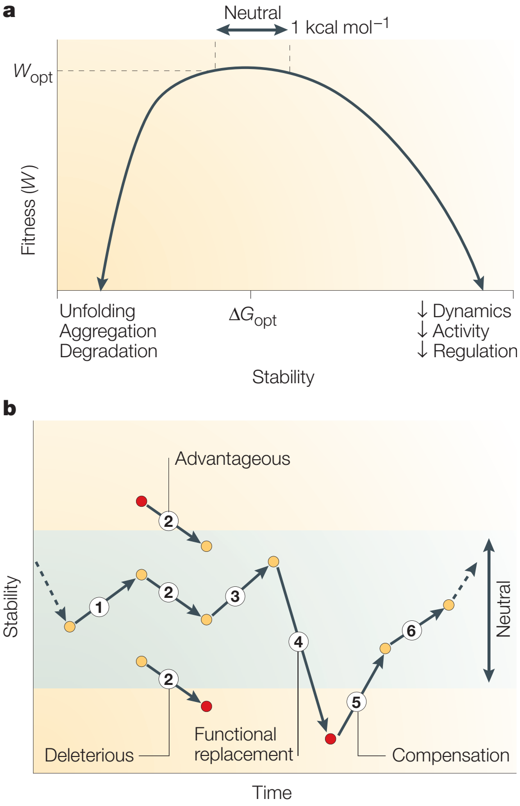

Figure 1 | **The relationship between protein stability, organismal fitness and protein evolution. a** | The relationship between stability ($\Delta G$) and fitness ($W$). The sharp decrease in fitness on the left is based on thermodynamic principles and observed effects of destabilizing mutations. The neutral zone of 1 kcal mol$^{-1}$ is based on the observation that proteins operate over a range of environmental temperatures. The decrease in fitness on the right is predicted to result from a reduction in function and increased aggregation owing to over-stabilization, leading to an inability to degrade proteins and control their expression. **b** | The evolution of protein stability as a constrained 'random walk' through sequence space. Protein sequences are represented as circles (yellow circles indicate sequences that are selectively neutral; red circles indicate those that have deleterious effects). Missense mutations are shown as the connecting labelled arrows. The series from 1 to 6 represents a trajectory of fixations through sequence space. The series of mutations from 1 to 3 represents a neutral 'meandering' through sequence space. The adaptive fixation 4, which is advantageous despite its effects on stability and aggregation, induces a strong selection pressure for the compensating mutation 5 to restore stability to the neutral zone. The three parallel occurrences of mutation 2 highlight the extensive epistasis that exists for fitness, as the same mutation can be advantageous, neutral or deleterious, depending on the current stability of the protein.

## A model for protein evolution

Here we present a model of protein evolution that integrates population genetic parameters such as organismal fitness and population size with our understanding of protein stability (BOX 1). This model focuses exclusively on stability because of the currently limited understanding of the fitness consequences of mutations that affect aggregation and degradation.

### Box 2 | Natural selection and mutational dynamics within populations

**The fate of new mutations**

After the occurrence of a new mutation, its fate is influenced both by the magnitude and sign of its selective effect and by stochastic processes that take place at the population-level[1], which are referred to as GENETIC DRIFT. For example, although it might be beneficial, a mutation might nevertheless be lost from a population while it is still rare, because random sampling of all the genotypes in a population takes place during biological reproduction. Formally, the probability of fixation of a novel mutation[1] is given in equation 1:

$$u(s,N) = \begin{cases} \dfrac{1-\exp(-2s)}{1-\exp(-4Ns)} & : s \neq 0 \\ \dfrac{1}{2N} & : s = 0 \end{cases} \qquad (1)$$

where $N$ is the effective population size and $s$ is the selection coefficient that represents the normalized fitness effect of the new mutation. The relative importance of genetic drift is inversely related to population size and, broadly speaking, drift dominates the process when $|Ns| < 1$. $u(s, N)$ is the probability of fixation function.

The evolution of a protein can conveniently be represented by its mutational trajectory through discrete sequence space[78,93], in which adjacent points represent pairs of sequences that differ by exactly one missense mutation. Projecting the fitness value of each sequence over this space defines a fitness landscape[78,81]. Except for the action of genetic drift, populations are traditionally regarded as evolving by the sequential fixation of individual beneficial mutations, represented by trajectories that never descend when plotted on this landscape[81].

**Population delocalization**

The picture of protein evolution described above is incomplete[94,95], because populations cannot be represented as a single point in sequence space in the presence of genetic variation[82,96]. This phenomenon is termed population delocalization and has important evolutionary implications. In particular, although the frequency of low-fitness protein sequences will be low, they might nevertheless acquire further mutations before elimination. If such a subsequent mutation yields a selectively compensated genotype, it can spread in the population by genetic drift[82,94] or selection[95,96], even though this mutational trajectory traverses a FITNESS VALLEY[81].

Populations can cross fitness valleys in one of two ways[96]. When populations are small relative to the depth of the valley (that is, $|Ns| < 1$), the rate of protein evolution is equal to the mutation rate because all missense mutations are effectively neutral[96]. As population size increases, the rate of protein evolution drops because the probability of fixation of deleterious mutation declines as natural selection becomes more effective (equation 1) (REF. 1.). However, populations now become delocalized and in this second regime the rate of protein evolution, although lower than in small populations, is again independent of population size[82,96]. See supplementary information S1,S3 (box and figure) for a full treatment of the population genetics principles that underlie these points.

Our stability model is based on two conclusions that are drawn from the preceding discussion. First, organismal fitness is a NON-MONOTONIC, concave function of protein stability, meaning that fitness decreases with increasing deviation from an optimal stability (BOX 1; FIG. 1a). Fitness decreases rapidly when $\Delta G$ departs from the neutral zone that is approximately 1 kcal mol$^{-1}$ wide. Outside this range, increased instability results in loss of activity and increased aggregation and degradation. Conversely, hyperstability leads to loss of flexibility and activity, increased aggregation and resistance to degradation. The model also incorporates a protein-specific parameter that reflects the maximum fitness cost of perturbing stability, which is determined by structural constraints on the protein and its role in the economy of the organism. This parameter links the biophysical consequences of mutations, which are broadly similar among proteins, to the contribution of individual proteins to organismal fitness, which varies widely among proteins.

The second conclusion from the biophysical studies that underlie our model is that most mutations alter protein stability, and the magnitude of this change is on the same order as the total protein stability. The distribution of mutational effects on stability is largely independent of the current $\Delta G$ of the protein. Although the precise details of this distribution are unclear, most mutations have a $\Delta\Delta G$ of between 0.5 and 5 kcal mol$^{-1}$ (regardless of the direction of the effect), with a minority having effects that are less than 0.5 kcal mol$^{-1}$ (REF. 45). Equation 2 in BOX 1 provides a mathematical representation of such a distribution.

Given the approximate additivity of mutations with respect to $\Delta G$, we can represent evolutionary trajectories through different stability values as shown in FIG. 1b. Successive mutations transform an initial $\Delta G$ through a series of steps through stability space. Selection constrains these trajectories to remain in or near the neutral range. One important aspect of this model is that fitness is a function of $\Delta G$, which implies that mutations do not have an intrinsic fitness effect, but rather that the fitness consequence of a mutation depends on the current $\Delta G$, which reflects the cumulative effect of all preceding fixation events.

So in this model there is strong epistasis — not with respect to $\Delta G$ itself, but as a consequence of the mapping of $\Delta G$ onto fitness. This epistasis for fitness arises

GENETIC DRIFT
The stochastic variation in population frequency of a mutation that is due to the sampling process inherent in reproduction.

FITNESS VALLEY
The circumstance in which mutations individually reduce fitness while jointly increasing it, so that when fitness is represented graphically, these single mutants form a valley.

NON-MONOTONIC
A function in which the first derivative changes sign. Here this indicates that fitness decreases with departure from an optimal stability.

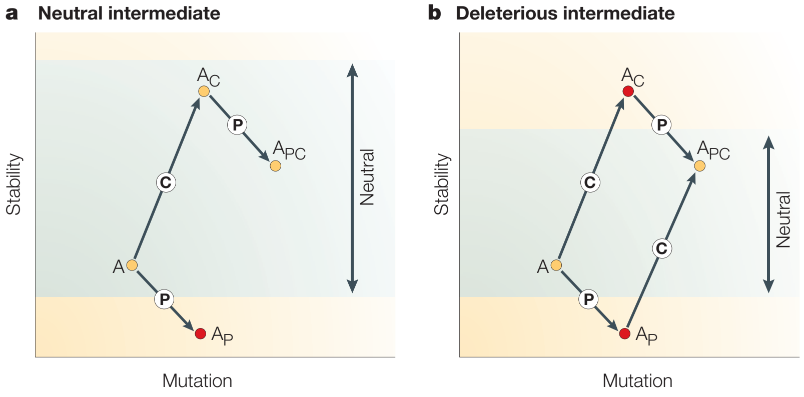

Figure 2 | **Two models for the occurrence of compensatory pathogenic deviations.** Two alternative explanations for how a pathogenic polymorphism (P) in protein sequence A can become fixed in a population. Protein sequences are represented as circles (yellow for neutral, red for deleterious). Missense mutations are shown as labelled arrows that connect the sequences. **a** | On its own, P reduces protein stability to a level that is outside the neutral zone and therefore cannot become fixed. However, an intermediate mutation C exists, which is neutral and can become fixed. C increases stability to a level that is still within the neutral zone and compensates for the subsequent reduction in stability that is caused by P, allowing P to become fixed in the sequence $A_{PC}$. **b** | Here C is also deleterious. However, population delocalization means that both $A_P$ and $A_C$ are present at low levels in the population. This allows the simultaneous fixation of both P and C, without fixing either of the intermediate low-fitness sequences $A_C$ or $A_P$.

from the non-monotonicity of the fitness function, and is a general property of STABILIZING SELECTION that acts on a trait even if mutations function additively on the trait itself. Our model shares broad similarities with LATTICE MODELS of protein evolution[79,80]. These models, however, assume that fitness is an increasing function of stability, whereas the evidence is for a non-monotonic function[36]. Non-monotonicity means that mutations can be advantageous on one background but neutral or deleterious on others (FIG. 1b) — a phenomenon known as sign epistasis[78], which has qualitatively different interpretations and implications for protein evolution.

Although simplified, our model provides a framework to evaluate the relationship between protein stability and evolution. As described below, it is consistent with many experimental observations. Furthermore, it makes predictions about the dynamics of protein evolution that are amenable to experimental or computational verification. It also motivates further experiments into new areas at the junction between biochemistry and evolutionary biology.

STABILIZING SELECTION
Selection that maintains a phenotype at some intermediate value.

LATTICE MODEL
An abstract model for protein folding in which a protein chain is constrained to occupy discrete points on a regular two- or three-dimensional lattice.

## Reconciling observations and theory

Here we re-examine the issues introduced in the first sections of this review in light of the model developed above. In particular we show how a model that integrates biophysical properties such as stability and aggregation can help us to understand a range of observations about protein evolution.

***Effects of missense mutations.*** The large $\Delta\Delta G$ of many missense mutations means that most single-hit mutations will produce proteins that have significantly deleterious $\Delta G$ values if the initial protein stability is near its optimum. For an essential protein with an optimum $\Delta G$ of –4 kcal mol$^{-1}$, >97% of mutations result in a fitness penalty of more than 1% according to the fitness and mutation distributions in BOX 1. According to the equations in BOX 2, this means that 3% of mutations behave neutrally in a population size of 100, and this figure falls to just 1% in a population of 100,000 individuals, indicating that the neutral dynamics are insensitive to population size.

***Implications for compensatory mutations.*** Functionally advantageous mutations — such as those that improve catalysis or change substrate specificity — often simultaneously perturb stability (FIG. 1b), generating a selective pressure to restore stability to the optimum $\Delta G$. For example, a mutation that increases the catalysis rate at the cost of 1 kcal mol$^{-1}$ of stability might increase overall organismal fitness by 10%, representing a functional advantage of 50% and a destabilization cost of 40%. For the parameters in BOX 1, when $\Delta\Delta G$ = 1 kcal mol$^{-1}$, 17% of the subsequent mutations would move stability closer to the optimum and therefore be beneficial. A more pertinent estimate for comparison with previous experiments[15,64,65,71,73] is to consider only mutations that restore stability to almost wild-type levels. In this case, only 4% of single compensatory mutants restore $\Delta G$ to within 0.25 kcal mol$^{-1}$ of $\Delta G_{opt}$.

The fraction of mutations that restore $\Delta G$ is predicted to increase with the deviation from $\Delta G_{opt}$. Continuing with our example, for a primary mutation with $\Delta\Delta G$ = 1.5 kcal mol$^{-1}$, this fraction is 29%, whereas if $\Delta\Delta G$ = 3 kcal mol$^{-1}$, 33% of secondary mutations are beneficial. Again, although these values are based on a simple model, it seems clear that there is a great deal of opportunity in nature for compensatory mutations to restore stability following a functionally advantageous fixation. One particularly important implication is that molecular adaptation will often occur through a cascade of missense mutations. Such behaviour would emerge as bursts of strongly selected missense fixations among phylogenetic lineages (for an example see REF. 3), a notably non-neutral process.

***Compensated pathogenic deviations.*** An observation that is closely related to the occurrence of compensatory mutations is that pathogenic missense mutations in one species are often found to be the wild-type state in orthologous proteins. This phenomenon, which relies on compensatory mutation, is known as compensated pathogenic deviation (CPD) and has important implications for protein evolution[9,70]. Our model, which is based on biophysical properties, also provides a framework for understanding this phenomenon, which can be accounted for by two possible hypotheses.

First, imagine that P is a pathogenic amino-acid replacement in species 1, whereas P is fixed in species 2. We have seen that P probably affects stability, so we can assume that P is destabilizing, although the following argument holds in reverse if P is stabilizing. If the wild-type sequence in species 1 is near the border of the neutral zone, then mutation to P in species 1 projects $\Delta G$ into the deleterious zone (FIG. 2a). However, in species 2, fixation of P is possible if the sequence already contains a compensatory mutation, C, that renders P neutral or beneficial. As P is destabilizing, a probable mechanism is for C to be stabilizing. This explains why P is deleterious in species 1, but not in the orthologous molecule in species 2. If C is neutral in species 2, then both mutations can be fixed without fitness cost. This model emphasizes the epistasis for stability that occurs with respect to fitness, in particular the fact that neutral substitutions can alter the opportunities for subsequent neutral mutation. Such interdependencies can profoundly affect the dynamics of neutral evolution (see the next section).

In a second hypothesis, if both P and C are individually deleterious but jointly selectively neutral (FIG. 2b), they give rise to a fitness valley[81] because their joint fixation seems to require a transient decline in fitness, which is an unlikely event. However, POPULATION DELOCALIZATION provides an alternative explanation for the emergence of such CPDs (BOX 2). Delocalization allows a population to simultaneously acquire both amino-acid replacements without fixing either deleterious intermediate. Importantly, this phenomenon occurs at a rate that is largely insensitive to population size[82], a point that is discussed further below. In the future, determining the stability and fitness of each mutation that comprises the CPD might differentiate between the two mechanisms illustrated in FIG. 2 by determining whether the compensatory mutation is neutral or deleterious in isolation.

*Implications for long-term patterns of protein evolution.* Models of protein evolution must also account for two fundamental and long-recognized facts about rates and patterns of divergence between species. First, the average rate of missense fixation for any given protein is approximately constant[1,83,84], even across species that are thought to represent a wide range of population sizes. This indicates the existence of a MOLECULAR CLOCK[83] and has been taken as strong evidence for the selective neutrality of missense fixation events[1] because the rate of fixation under the neutral model is independent of population size. By contrast, because the efficacy of natural selection is inversely related to population size (BOX 2), models that invoke the action of selection in the fixation of missense mutations predict a positive relationship between their rate of fixation and population size. Second, the variance in the rate of missense fixation across lineages significantly exceeds its mean[5,84–86] (described as 'overdispersion' of the molecular clock), a fact that is at odds with expectations under the neutral model in which the variance and mean should be equal[5]. So far it has proved difficult to theoretically account for these observations[87].

Our model indicates how these facts might be reconciled. First, both hypotheses outlined in FIG. 2 predict a rate of missense fixation that is largely independent of population size. In the first case (FIG. 2a) both missense fixations are selectively neutral and the rate of fixation is proportional only to the mutation rate. In the second case (FIG. 2b) missense mutations are fixed in pairs that are jointly neutral. Here the rate of missense fixation is proportional to the product of the deleterious and compensatory mutation rates, independent of population size (BOX 2; see also supplementary information S4 (figure)). The many targets for mutation that affect protein stability indicate that the rate of fixation by this process might be substantially larger than previously appreciated[88].

Our model also offers an interpretation for the overdispersion of the molecular clock. Under the hypothesis depicted in FIG. 2a, the fixation of a neutral mutation intrinsically changes the availability of other neutral mutations, which can increase the dispersion index by altering the neutral mutation rate in that lineage[80]. In FIG. 2b, multiple missense mutations are likely to be fixed simultaneously in large populations. This causes a proportional increase in the mean number of fixations among lineages, but importantly also causes a squared increase in the variance in this quantity, resulting in a variance-to-mean ratio that is greater than unity. Therefore, it seems that some properties of the molecular clock might arise from compensatory evolution within protein stability, which is in part mediated by population effects such as delocalization.

*Implications for tests of neutrality.* As nucleotide data have accumulated, another line of evidence has further challenged the neutral interpretation of missense fixation events, which also predicts that patterns of molecular polymorphism within a population should mimic patterns of divergence between species[89]. Contrary to this, recent analyses indicate that between 45% (REF. 90) and 94% (REF. 4) of all recent missense fixations among species are due to natural selection.

Under our model, an advantageous functional fixation will often require subsequent fixations to restore protein stability and other biophysical properties, which would result in bursts of multiple missense fixations. Moreover, the process of population delocalization might elevate the number of missense fixations above expectations that are based on levels of polymorphism. Therefore, the substantial evidence for positive natural selection comes as no surprise in light of the above model. We emphasize, however, a point made elsewhere[91] that such positive selection might not signal increased adaptation. Indeed, our model highlights an evolutionary process in which positive selection is required only to maintain the *status quo*. Evolving protein sequences are not necessarily becoming 'better', although they are becoming different.

## Limitations, predictions and future work

Although our model considers only stability, we believe that other biophysical properties could be incorporated into an extended, multi-dimensional model. Aggregation, folding kinetics and native-state dynamics are especially interesting and important. Furthermore, the interplay between the complexity of functional adaptation[92], fluctuating environments[5] and stabilizing selection has not been adequately explored.

POPULATION DELOCALIZATION
A mechanism by which large populations can traverse fitness valleys without the fixation of deleterious mutational intermediates.

MOLECULAR CLOCK
The constant (clock-like) rate of missense fixation over evolutionary timescales.

Our model is most clearly applicable to two-state folding enzymes and it remains an open question whether it is informative about proteins that have a purely structural role, are membrane-bound, fold with multi-state behaviour, interact primarily with other proteins or are even natively unfolded.

Much further work is needed to fully develop and explore the model of protein evolution presented here. Critical parameters of the model lack experimental detail, although they can be estimated using straightforward, if laborious, technologies. First, the true relationship between organismal fitness and protein stability and other biophysical properties should be determined, perhaps through extensive mutagenesis experiments and growth rate, stability and aggregation assays in bacteria. Another way of determining this relationship would be to obtain $\Delta G$ values and aggregation rates for a set of orthologous proteins from species that live at equivalent environmental temperatures. In addition, the $\Delta G$ and $k_{agg}$ values of a large and unbiased library of mutant proteins should be determined to estimate the distribution of mutational effects on stability and aggregation. Ideally, the variability in these distributions should be assessed using a range of test proteins.

Despite the need for further work, our simplified model makes many interesting predictions that are amenable to experimental verification. One clear prediction is that, because protein stabilities fall within a tight range, natural selection disfavours highly stable proteins. We also expect to see a relationship between the magnitude of the stability perturbation owing to a functional mutation and the number of compensatory mutations. Finally, the inverse relationship between population size ($N$) and the strength of natural selection (BOX 2) implies sensitivity to $N$ in the distribution of stability and aggregation seen in standing genetic variation. Such relationships are becoming testable and will help to clarify the role that biophysical properties have in protein evolution.

## Conclusions

The relevance of protein biophysics to evolutionary problems has been generally unappreciated among evolutionary geneticists. Conversely, central aspects of population genetics — such as the importance of stochasticity and organismal fitness in natural selection — have been absent from the biochemical approach to protein evolution. In this review we have advanced a population-based model of protein evolution that is based on fundamental biophysical properties such as stability and aggregation. This model explains many observations about protein evolution, both old and new, and offers a framework for future investigations. We hope that it will allow the exchange of ideas and outstanding problems between evolutionary geneticists and biochemists to advance our understanding of the forces and constraints involved in protein evolution.

1. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
2. Blundell, T. L. & Wood, S. P. Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature* **257**, 197–203 (1975).
**The authors present the fundamental biochemical argument against the neutral theory of evolution.**
3. Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S. & Kondrashov, A. S. Positive selection at sites of multiple amino acid replacements since rat–mouse divergence. *Nature* **429**, 558–562 (2004).
4. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (Suppl. 1), 154–164 (2003).
5. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
**This is a fantastic discussion of the problems of protein evolution from an eminent population geneticist.**
6. Poon, A. & Chao, L. The rate of compensatory mutation in the DNA bacteriophage φX174. *Genetics* 23 May 2005 (doi:10.1534/genetics.104.039438).
7. Poon, A., Davis, B. H. & Chao, L. The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. *Genetics* 6 May 2005 (doi:10.1534/genetics.104.037259).
8. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.* **320**, 85–95 (2002).
**This paper describes the role of biochemistry in the evolution of antibiotic resistance genes.**
9. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99**, 14878–14883 (2002).
10. Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
**This is an introduction to protein aggregation and disease.**
11. Beadle, B. M. & Shoichet, B. K. Structural bases of stability–function tradeoffs in enzymes. *J. Mol. Biol.* **321**, 285–296 (2002).
12. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. A relationship between protein stability and protein function. *Proc. Natl Acad. Sci. USA* **92**, 452–456 (1995).
13. Bull, J. J., Badgett, M. R. & Wichman, H. A. Big-benefit mutations in a bacteriophage inhibited with heat. *Mol. Biol. Evol.* **17**, 942–950 (2000).
14. Wilson, K. P., Malcolm, B. A. & Matthews, B. W. Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme. *J. Biol. Chem.* **267**, 10842–10849 (1992).
15. Mitraki, A., Fane, B., Haase-Pettingell, C., Sturtevant, J. & King, J. Global suppression of protein folding defects and inclusion body formation. *Science* **253**, 54–58 (1991).
16. Zwanzig, R. Two-state models of protein folding kinetics. *Proc. Natl Acad. Sci. USA* **94**, 148–150 (1997).
17. Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochem.* **39**, 11177–11183 (2000).
18. Pace, C. N. The stability of globular proteins. *CRC Crit. Rev. Biochem.* **3**, 1–43 (1975).
19. Creighton, T. E. *Proteins: Structures and Molecular Properties* (W. H. Freeman and Company, New York, 1993).
20. Chiti, F. *et al.* Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J.* **19**, 1441–1449 (2000).
21. van den Berg, B., Ellis, R. J. & Dobson, C. M. Effects of macromolecular crowding on protein folding and aggregation. *EMBO J.* **18**, 6927–6933 (1999).
22. Pawar, A. P. *et al.* Prediction of 'aggregation-prone' and 'aggregation-susceptible' regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.* **350**, 379–392 (2005).
23. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424**, 805–808 (2003).
24. Chiti, F. *et al.* Kinetic partitioning of protein folding and aggregation. *Nature Struct. Biol.* **9**, 137–143 (2002).
25. Ramirez-Alvarado, M., Merkel, J. S. & Regan, L. A systematic exploration of the influence of the protein stability on amyloid fibril formation *in vitro*. *Proc. Natl Acad. Sci. USA* **97**, 8979–8984 (2000).
26. Bucciantini, M. *et al.* Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**, 507–511 (2002).
27. Hartl, F. U. & Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**, 1852–1858 (2002).
28. Georgiou, G., Valax, P., Ostermeier, M. & Horowitz, P. M. Folding and aggregation of TEM β-lactamase: analogies with the formation of inclusion bodies in *Escherichia coli*. *Protein Sci.* **3**, 1953–1960 (1994).
29. Glickman, M. H. & Ciechanover, A. The ubiquitin–proteasome proteolytic pathway: destruction for the sake of construction. *Phys. Rev.* **82**, 373–428 (2002).
30. Parsell, D. A. & Sauer, R. T. The structural stability of a protein is an important determinant of its proteolytic susceptibility in *Escherichia coli*. *J. Biol. Chem.* **264**, 7590–7595 (1989).
31. Goldberg, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895–899 (2003).
**The author provides an introduction to protein degradation and disease.**
32. Gregersen, N., Bross, P., Jorgensen, M. M., Corydon, T. J. & Andresen, B. S. Defective folding and rapid degradation of mutant proteins is a common disease mechanism in genetic disorders. *J. Inherit. Metab. Dis.* **23**, 441–447 (2000).
33. Pakula, A. A. & Sauer, R. T. Genetic analysis of protein stability and function. *Annu. Rev. Genet.* **23**, 289–310 (1989).
**This is an excellent review of mutational effects on protein stability.**
34. Fields, P. A. Review: Protein function at thermal extremes: balancing stability and flexibility. *Comp. Biochem. Physiol. A* **129**, 417–431 (2001).
35. Daniel, R. M., Dunn, R. V., Finney, J. L. & Smith, J. C. The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 69–92 (2003).
36. Somero, G. N. Proteins and temperature. *Annu. Rev. Physio.* **57**, 43–68 (1995).
**This paper describes the adaptation of proteins to environmental temperature.**
37. Fink, A. L. Natively unfolded proteins. *Curr. Opin. Struct. Biol.* **15**, 35–41 (2005).

38. Ferrer, M., Chernikova, T. N., Yakimov, M. M., Golyshin, P. N. & Timmis, K. N. Chaperonins govern growth of *Escherichia coli* at low temperatures. *Nature Biotechnol.* **21**, 1266–1267 (2003).
39. Daopin, S., Alber, T., Baase, W. A., Wozniak, J. A. & Matthews, B. W. Structural and thermodynamic analysis of the packing of two $\alpha$-helices in bacteriophage T4 lysozyme. *J. Mol. Biol.* **221**, 647–667 (1991).
40. Green, S. M. & Shortle, D. Patterns of nonadditivity between pairs of stability mutations in staphylococcal nuclease. *Biochem.* **32**, 10131–10139 (1993).
41. Matthews, B. W. Studies on protein stability with T4 lysozyme. *Adv. Protein Chem.* **46**, 249–278 (1995).
42. Milla, M. E., Brown, B. M. & Sauer, R. T. Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nature Struct. Biol.* **1**, 518–523 (1994).
43. Pakula, A. A. & Sauer, R. T. Amino acid substitutions that increase the thermal stability of the $\lambda$Cro protein. *Proteins* **5**, 202–210 (1989).
44. Shortle, D. Probing the determinants of protein folding and stability with amino acid substitutions. *J. Biol. Chem.* **264**, 5315–5318 (1989).
45. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
46. Fersht, A. R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771–782 (1992).
47. Alber, T. Mutational effects on protein stability. *Annu. Rev. Biochem.* **58**, 765–798 (1989).
48. Schultz, S. C. & Richards, J. H. Site-saturation studies of $\beta$-lactamase: production and characterization of mutant $\beta$-lactamases with all possible amino acid substitutions at residue 71. *Proc. Natl Acad. Sci. USA* **83**, 1588–1592 (1986).
49. Pakula, A. A., Young, V. B. & Sauer, R. T. Bacteriophage $\lambda$*cro* mutations: effects on activity and intracellular degradation. *Proc. Natl Acad. Sci. USA* **83**, 8829–8833 (1986).
50. Rosen, R. *et al.* Protein aggregation in *Escherichia coli*: role of proteases. *FEMS Microbiol. Lett.* **207**, 9–12 (2002).
51. Calloni, G., Zoffoli, S., Stefani, M., Dobson, C. M. & Chiti, F. Investigating the effects of mutations on protein aggregation in the cell. *J. Biol. Chem.* **280**, 10607–10613 (2005).
52. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnol.* **22**, 1302–1306 (2004).
53. Broome, B. M. & Hecht, M. H. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J. Mol. Biol.* **296**, 961–968 (2000).
54. Ventura, S. *et al.* Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl Acad. Sci. USA.* **101**, 7258–7263 (2004).
55. Zavodszky, P., Kardos, J., Svingor, A. & Petsko, G. A. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl Acad. Sci. USA* **95**, 7406–7411 (1998).
56. Fersht, A. R. *Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).
57. Kelly, J. W. & Balch, W. E. Amyloid as a natural product. *J. Cell Biol.* **161**, 461–462 (2003).
58. Perl, D., Mueller, U., Heinemann, U. & Schmid, F. X. Two exposed amino acid residues confer thermostability on a cold shock protein. *Nature Struct. Biol.* **7**, 380–383 (2000).
**This article demonstrates the stabilizing effect of mutations during thermoadaptation.**
59. Perl, D. & Schmid, F. X. Some like it hot: the molecular determinants of protein thermostability. *Chembiochem* **3**, 39–44 (2002).
60. Wagner, G. P. & Gabriel, W. Quantitative variation in finite parthenogenetic populations: what stops Muller's ratchet in the absence of recombination? *Evolution* **44**, 715–731 (1990).
61. Schrag, S. J., Perrot, V. & Levin, B. R. Adaptation to the fitness costs of antibiotic resistance in *Escherichia coli*. *Proc. R. Soc. Lond. B* **264**, 1287–1291 (1997).
62. Maisnier-Patin, S., Berg, O. G., Liljas, L. & Andersson, D. I. Compensatory adaptation to the deleterious effect of antibiotic resistance in *Salmonella typhimurium*. *Mol. Microbiol.* **46**, 355–366 (2002).
**This paper describes the fitness costs of streptomycin resistance and the compensatory mutations that are involved.**
63. Poteete, A. R., Rennell, D., Bouvier, S. E. & Hardy, L. W. Alteration of T4 lysozyme structure by second-site reversion of deleterious mutations. *Protein Sci.* **6**, 2418–2425 (1997).
64. Shortle, D. & Lin, B. Genetic analysis of staphylococcal nuclease: identification of three intragenic 'global' suppressors of nuclease-minus mutations. *Genetics* **110**, 539–555 (1985).
65. Mitraki, A., Danner, M., King, J. & Seckler, R. Temperature-sensitive mutations and second-site suppressor substitutions affect folding of the P22 tailspike protein *in vitro*. *J. Biol. Chem.* **268**, 20071–20075 (1993).
66. Levin, B. R., Perrot, V. & Walker, N. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* **154**, 985–997 (2000).
67. Nagaev, I., Bjorkman, J., Andersson, D. I. & Hughes, D. Biological cost and compensatory evolution in fusidic acid-resistant *Staphylococcus aureus*. *Mol. Microbiol.* **40**, 433–439 (2001).
68. Bjorkman, J., Hughes, D. & Andersson, D. I. Virulence of antibiotic-resistant *Salmonella typhimurium*. *Proc. Natl Acad. Sci. USA* **95**, 3949–3953 (1998).
69. Burch, C. L. & Chao, L. Evolution by small steps and rugged landscapes in the RNA virus $\phi$6. *Genetics* **151**, 921–927 (1999).
70. Kulathinal, R. J., Bettencourt, B. R. & Hartl, D. L. Compensated deleterious mutations in insect genomes. *Science* **306**, 1553–1554 (2004).
71. Sideraki, V., Huang, W., Palzkill, T. & Gilbert, H. F. A secondary drug resistance mutation of TEM-1 $\beta$-lactamase that suppresses misfolding and aggregation. *Proc. Natl Acad. Sci. USA* **98**, 283–288 (2001).
72. Kim, H. W. *et al.* Restoring allosterism with compensatory mutations in hemoglobin. *Proc. Natl Acad. Sci. USA* **91**, 11547–11551 (1994).
73. Mateu, M. G. & Fersht, A. R. Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to impaired hetero-oligomerization. *Proc. Natl Acad. Sci. USA* **96**, 3595–3599 (1999).
74. Andersson, D. I. & Levin, B. R. The biological cost of antibiotic resistance. *Curr. Opin. Microbiol.* **2**, 489–493 (1999).
75. Borman, A. M., Paulous, S. & Clavel, F. Resistance of human immunodeficiency virus type 1 to protease inhibitors: selection of resistance mutations in the presence and absence of the drug. *J. Gen. Virol.* **77**, 419–426 (1996).
76. Rutherford, S. L. Between genotype and phenotype: protein chaperones and evolvability. *Nature Rev. Genet.* **4**, 263–274 (2003).
77. Sangster, T. A., Lindquist, S. & Queitsch, C. Under cover: causes, effects and implications of Hsp90-mediated genetic capacitance. *Bioessays* **26**, 348–362 (2004).
78. Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**, 1165–1174 (2005).
79. Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA* **102**, 606–611 (2005).
80. Bastolla, U., Roman, H. E. & Vendruscolo, M. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.* **200**, 49–64 (1999).
81. Wright, S. in *Proc. 6th Int. Congr. Genet.* (ed. Jones, D. F.) 356–366 (Brooklyn Botanic Garden, Menasha, Wisconsin, 1932).
82. Stephan, W. The rate of compensatory evolution. *Genetics* **144**, 419–426 (1996).
83. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins, a Symposium* (eds Bryson, V. & Vogel, H.) 97–166 (Academic Press, New York, 1965).
84. Wilson, A. C., Carlson, S. S. & White, T. J. Biochemical evolution. *Annu. Rev. Biochem.* **46**, 573–639 (1977).
85. Ohta, T. & Kimura, M. On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1**, 18–25 (1971).
86. Langley, C. H. & Fitch, W. M. An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3**, 162–177 (1974).
87. Cutler, D. J. Understanding the overdispersed molecular clock. *Genetics* **154**, 1403–1417 (2000).
88. Gillespie, J. H. Molecular evolution over the mutational landscape. *Evolution* **38**, 1116–1129 (1984).
**This is the first rigorous treatment of sequence evolution through complex fitness landscapes.**
89. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
90. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
91. Orr, A. H. The genetic theory of adaptation: a brief history. *Nature Rev. Genet.* **6**, 119–127 (2005).
92. Hartl, D. L., Dykhuizen, D. E. & Dean, A. M. Limits of adaptation: the evolution of selective neutrality. *Genetics* **111**, 655–674 (1985).
93. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* **225**, 563–565 (1974).
94. Kimura, M. The role of compensatory neutral mutations in molecular evolution. *J. Genet.* **64**, 7–19 (1985).
95. Carter, A. J. R. & Wagner, G. P. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. *Proc. R. Soc. Lond. B* **269**, 953–960 (2002).
96. Weinreich, D. M. & Chao, L. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution* **59**, 1175–1182 (2005).
97. Bross, P. *et al.* Protein misfolding and degradation in genetic diseases. *Human Mut.* **14**, 186–198 (1999).
98. Pedersen, C. B. *et al.* Misfolding, degradation, and aggregation of variant proteins. The molecular pathogenesis of short chain acyl-CoA dehydrogenase (SCAD) deficiency. *J. Biol. Chem.* **278**, 47449–47458 (2003).
99. Haass, C. & Steiner, H. Alzheimer disease $\gamma$-secretase: a complex story of GxGD-type presenilin proteases. *Trends Cell Biol.* **12**, 556–562 (2002).
100. Aguzzi, A. & Haass, C. Games played by rogue proteins in prion disorders and Alzheimer's disease. *Science* **302**, 814–818 (2003).
101. Sherman, M. Y. & Goldberg, A. L. Cellular defenses against unfolded proteins: a cell biologist thinks about neurodegenerative diseases. *Neuron* **29**, 15–32 (2001).
102. Venkatraman, P., Wetzel, R., Tanaka, M., Nukina, N. & Goldberg, A. L. Eukaryotic proteasomes cannot digest polyglutamine sequences and release them during degradation of polyglutamine-containing proteins. *Mol. Cell* **14**, 95–104 (2004).
103. Eaton, W. A. & Hofrichter, J. Sickle cell hemoglobin polymerization. *Adv. Protein Chem.* **40**, 63–279 (1990).

## Acknowledgements

M.A.D. is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation. D.M.W. and D.L.H. thank the National Science Foundation for support.

## Competing interests statement

The authors declare no competing financial interests.

## Online links

**DATABASES**

**The following terms in this article are linked online to:**

**OMIM:** http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Alzheimer disease | Huntington disease

**SUPPLEMENTARY INFORMATION**

See online article: S1 (box) | S2 (figure) | S3 (figure) | S4 (figure)

**Access to this links box is available online.**